Case Studies 2022L

Model performance measures

Apr 7, 2022

Usage purposes of measure

- model evaluation
- model comparison
- out-of-sample comparisons

MPMs for regression models

Assume that the train set with n observations on p explanatory variables and on a dependent variable Y. Let x_i denote the vector of values of the explanatory variables for the i-th observation, and y_i is the corresponding value of the dependent variable.

The train data is used to train model f(.) and $\hat{y}_i$ is used to denote the prediction for y_i

Mean-squared error (MSE)

The most popular measure for regression models is the mean-squared error, defined as

$$MSE(f, X, y) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2$$

MSE weights all differences (r_i) equally, so large residuals have got a high impact on MSE. Thus, it is sensitive to outliers.

For a perfect model, which predicts all y_i exactly, MSE = 0.

Root-Mean-squared error (RMSE)

MSE is constructed on a different scale from the dependent variable. Thus, a more interpretable variant of this measure is the root-mean-squared-error (RMSE), defined as

$$RMSE(f, X, y) = \sqrt{MSE(f, X, y)}$$

R^2

R^2 (coefficient of determination) is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

$$R^2(f, X, y) = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

In the best case, the predicted values exactly match the observed values, which results in the numerator will be equal to zero and R^2 = 1. A baseline model, which always predicts as the mean of y, will have R^2 = 0.

Median absolute-deviation (MAD)

Given sensitivity of MSE to outliers, sometimes the median absolute-deviation (MAD) is considered:

$$MAD(f, X, y) = median(|r_1|, \ldots, |r_n|)$$

MPMs for classification models

In classification problems, the comparison of the observed and predicted values of the dependent variable for the n observations in a dataset can be summarized in **confusion table/matrix:**

	True value: `success`	**True value:** `failure`	**Total**
$\hat{y}_i \geq C$, predicted: `success`	True Positive: TP_C	False Positive (Type I error): FP_C	P_C
$\hat{y}_i < C$, predicted: `failure`	False Negative (Type II error): FN_C	True Negative: TN_C	N_C
Total	S	F	n

To predict the label of the observations, the predicted probability is compared to a fixed cut-off threshold C (usually C = 0.5).

MPMs for classification models

Accuracy is the fraction of correct predictions in the entire testing dataset.

$$Accuracy = \frac{TP_C + TN_C}{n}$$

Precision is referred to as the *positive predictive value*.

$$Precision = \frac{TP_C}{TP_C + FP_C} = \frac{TP_C}{P_C}$$

Recall is the fraction of correct predictions among the true successes. It is also called **sensitivity.**

$$Recall = \frac{TP_C}{TP_C + FN_C} = \frac{TP_C}{S}$$

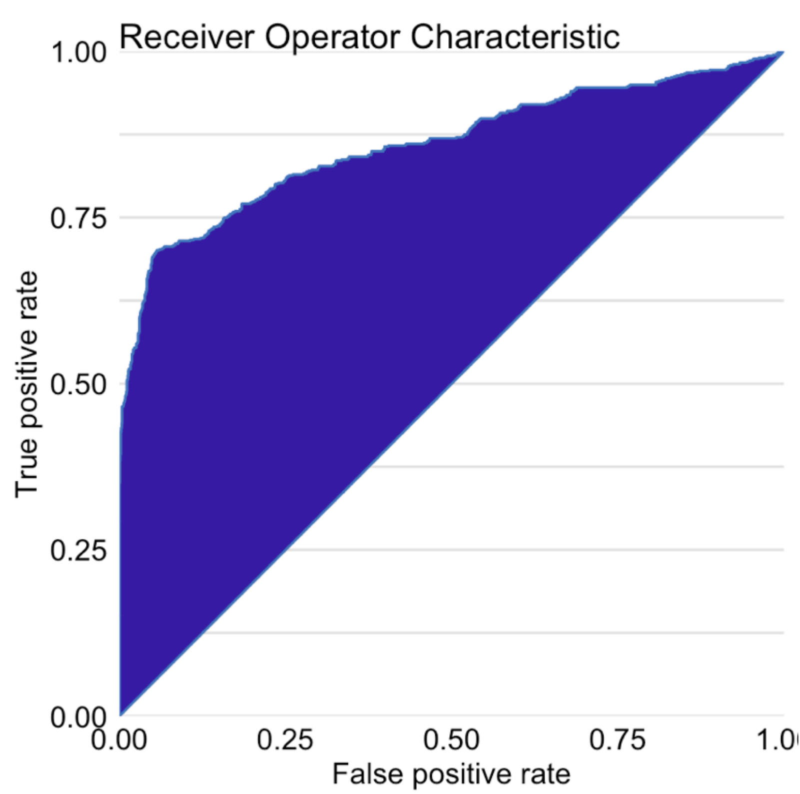
Specificity is referred to as the *true-negative rate*.

$$Specificity = \frac{TN_C}{TN_C + FP_C} = \frac{TN_C}{F}$$

ROC Curve

All measures depend on the choice of cut-off value is C. To assess the form and the strength of dependence, a common approach is to construct the **Receiver Operating Characteristic (ROC)** curve.

The ROC curve plots $Sensitivity_C$ in function of $1 - Sensitivity_C$ for all possible, ordered values of *C*.



There is a need to summarize the ROC curve with one number, which can be used to compare models. A popular measure that is used toward this aim is the **area under the curve (AUC)**.

Caution!

In the classification tasks,

If the dataset is imbalanced means the number of the levels of the target variable are quite different, the accuracy may be misleading result. In that case, you should check the model performance also by using the balanced accuracy:

$$BalancedAccuracy = \frac{1}{2}\left(\frac{TP_C}{TP_C + FN_C} + \frac{TN_C}{TN_C + FP_C}\right)$$

Please feel free to send e-mail about your questions!

📩 mustafa.cavus@pw.edu.pl